Perla Molina
BUS 315
Dadgar
December 06, 2022

Performing Data Mining Methods on Ovarian Cancer Data

I. INTRODUCTION

I've decided to work on Ovarian Cancer data from Kaggle. The data can be accessed through this link: https://www.kaggle.com/datasets/saurabhshahane/predict-ovarian-cancer. This link has multiple datasets explained on the webpage. I will be using the complete raw data with 51 variables and 349 observations. I will have to remain the file to 'ovarian.xlsx' and use that for the assignment.

At first glance, this dataset contains full of numerical values with abbreviations. The variables consist of complex biological aspects of the human body related to ovarian cancer, as well as biological features of the female sex body and the tumors. All variables are listed below with what the abbreviations mean:

- SUBJECT_ID
- AFP (alpha-fetoprotein)
- AG (Anion gap)
- Age
- ALB (albumin)
- ALP (Alkaline phosphatase)
- ALT (Alanine aminotransferase)
- AST (Aspartate aminotransferase)
- BASO# or BASO. (Basophil Cell Count)
- BASO% or BASO..1 (Basophil Cell ratio)
- BUN (blood urea nitrogen)
- Ca (calcium)
- CA125 (Carbohydrate antigen 125)
- CA19-9 (Carbohydrate antigen 19-9)
- CA72-4 (Carbohydrate antigen 72-4)
- CEA (Carcinoembryonic antigen)
- CL (chlorine)
- CO2CP (carbon dioxide-combining Power)
- CREA (creatinine)
- TYPE (1 = BOT for Benign Ovarian Tumor and 0 = OC for Ovarian Cancer)
- DBIL (direct bilirubin)
- EO# or EO. (eosinophil count)
- EO% or EO..1 (eosinophil ratio)
- GGT (Gama glutamyl transferase)
- GLO (globulin)
- GLU. (glucose)
- HCT (hematocrit)
- HE4 (human epididymis protein 4)
- HGB (hemoglobin)
- IBIL (Indirect bilirubin)
- K (kalium)
- LYM# or LYM. (lymphocyte count)

- LYM% or LYM..1 (lymphocyte ratio)
- MCH (Mean corpuscular hemoglobin)
- MCV (mean corpuscular volume)
- Menopause (0 for no, 1 for yes)
- Mg (magnesium)
- MONO# or MONO. (mononuclear cell count)
- MONO% or MONO..1 (monocyte ratio)
- MPV (Mean platelet volume)
- Na (Natrium)
- NEU (neutrophil ratio)
- PCT (thrombocytocrit)
- PDW (Platelet distribution width)
- PHOS (phosphorus)
- PLT (platelet count)
- RBC (Red blood cell count)
- RDW (red blood cell distribution width)
- TBIL (total bilirubin)
- TP (Total protein)
- UA (urine acid)

II. INVESTIGATE THE DATA

To begin investigating the data, I will first look at the summary to insure that the variable types are consistent and conduct any necessary cleaning. I will also use the str() function to check for inconsistent variable types. I already see that some variables are labeled as strings and will need to change that. The variables that I had to change from strings to numeric were AFP, CA125, and CA19.9. I believe this error when I loaded the dataset is due to how the .csv file was set up. I will also have to change the variables TYPE and Menopause into factors because TYPE indicates what kind of ovarian cancer the patient has and Menopause is a yes or no response to if the patient is under menopause.

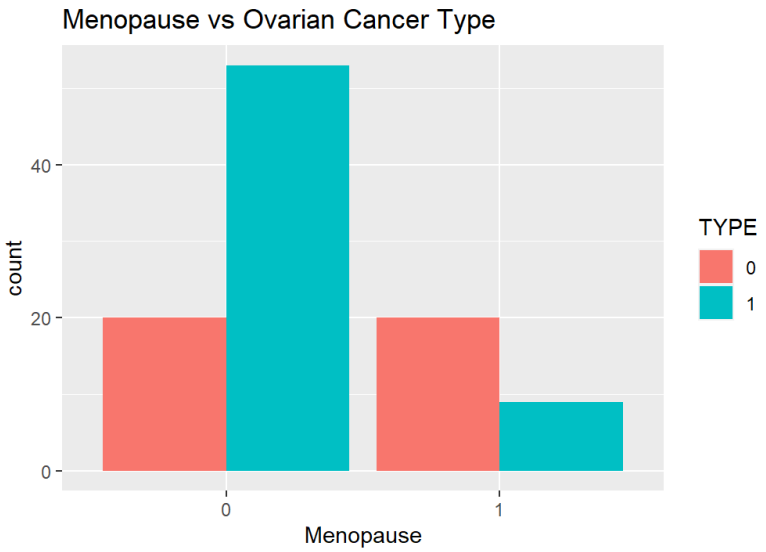
```
'data.frame':   349 obs. of  51 variables:
 $ SUBJECT_ID: int  1 2 3 4 5 6 7 8 9 10 ...
 $ AFP       : chr  "3.58\t" "34.24\t" "1.50\t" "2.75" ...
 $ AG        : num  19.4 24 18.4 16.6 20 ...
 $ Age       : int  47 61 39 45 45 44 53 76 38 30 ...
 $ ALB       : num  45.4 39.9 45.4 39.2 35 32.9 NA 50.4 36.3 40.8 ...
 $ ALP       : int  56 95 77 26 47 118 NA 76 64 77 ...
 $ ALT       : int  11 9 9 16 21 51 NA 16 62 16 ...
 $ AST       : int  24 13 18 17 27 32 NA 23 47 12 ...
 $ BASO.     : num  0.01 0.02 0.03 0.05 0.01 0.02 0.02 0.04 0 0.08 ...
 $ BASO..1   : num  0.3 0.3 0.6 0.74 0.1 0.42 0.5 0.5 0.4 1.23 ...
 $ BUN       : num  5.35 3.21 3.8 5.27 4.89 4.47 2.6 5.05 2.5 5.78 ...
 $ Ca        : num  2.48 2.62 2.57 2.35 2.48 2.49 2.24 2.68 2.42 2.43 .
 $ CA125     : chr  "15.36\t" "2444.00\t" "56.08\t" "2555" ...
```

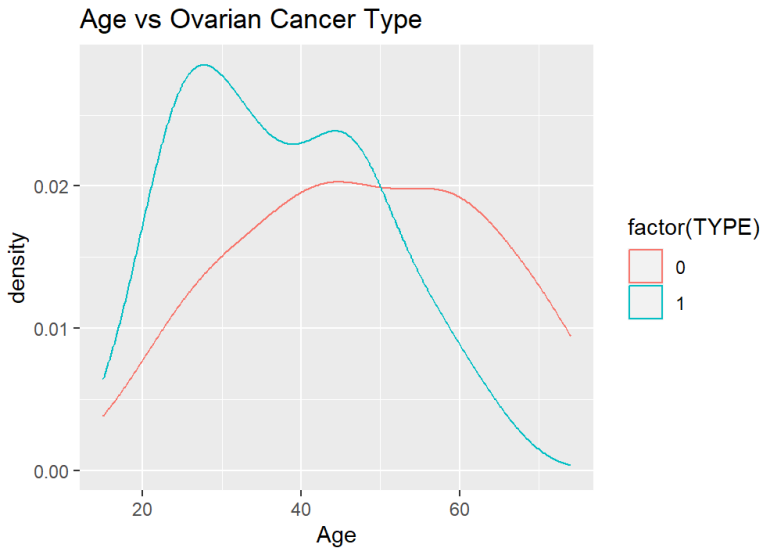
Now I will check and remove any NAs in the dataset to ensure it is completely cleaned. After removing the NAs, I checked the dimensions to see how many observations we are left with and it cut down to 102 observations.

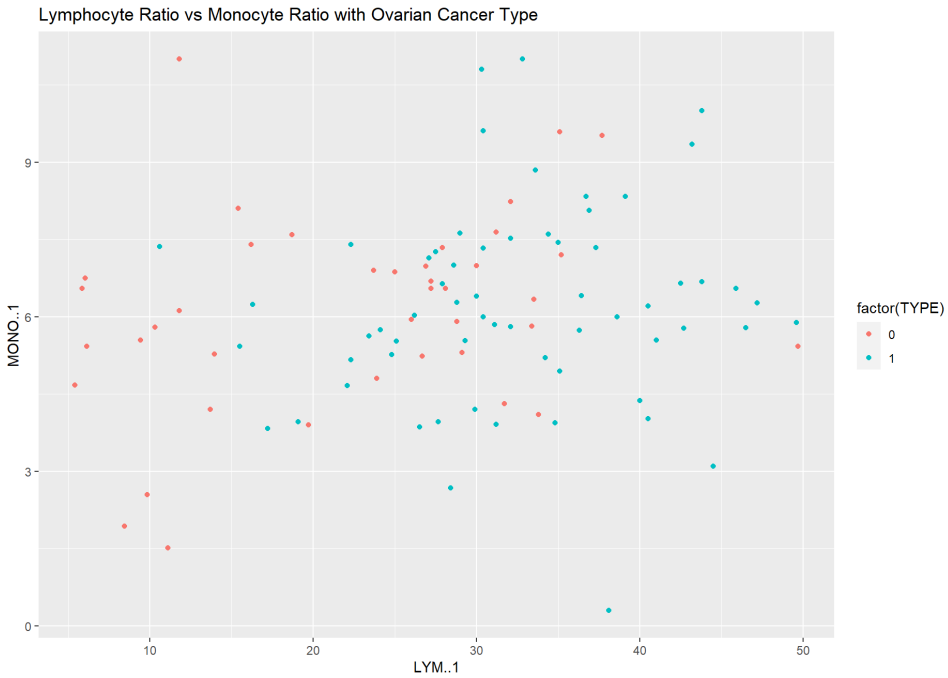
Next, I will look at some basic visuals of the demographics of our data. I want to see the proportions of patients in menopause and their cancer types to visually see if there is a correlation between those two categories. Looking at the bar graphic, there is a clear distinction

between having Benign Ovarian Tumors (TYPE = 1) and not experiencing menopause. Additionally, it seems as if having Ovarian Cancer (TYPE = 0) has relatively no effect on undergoing menopause.

I will also look at a graph of Age and cancer TYPE to also see if there is a correlation between age and acquiring ovarian cancer. Looking at a density graph of this, there is somewhat of a correlation. There is a significant amount of patients in their 20s and 30s with Benign Ovarian Cancer (TYPE = 1). This can make sense since Ovarian Tumors can be relatively benign when the patient is younger and depending on the circumstances, it could get worse over time. Similarly, there seem to be more older patients with Ovarian Cancer (TYPE = 0). This makes sense, healthcare-wise because most patients with Ovarian Cancer tend to be over the age of 50 since it is a type of cancer that is acquired over time. It is also due to the lack of available diagnosis technology in this niche of cancer as there are no current forms of early detection and thus, Ovarian Cancer oftentimes is detected in the late stages of life and tumor growth.

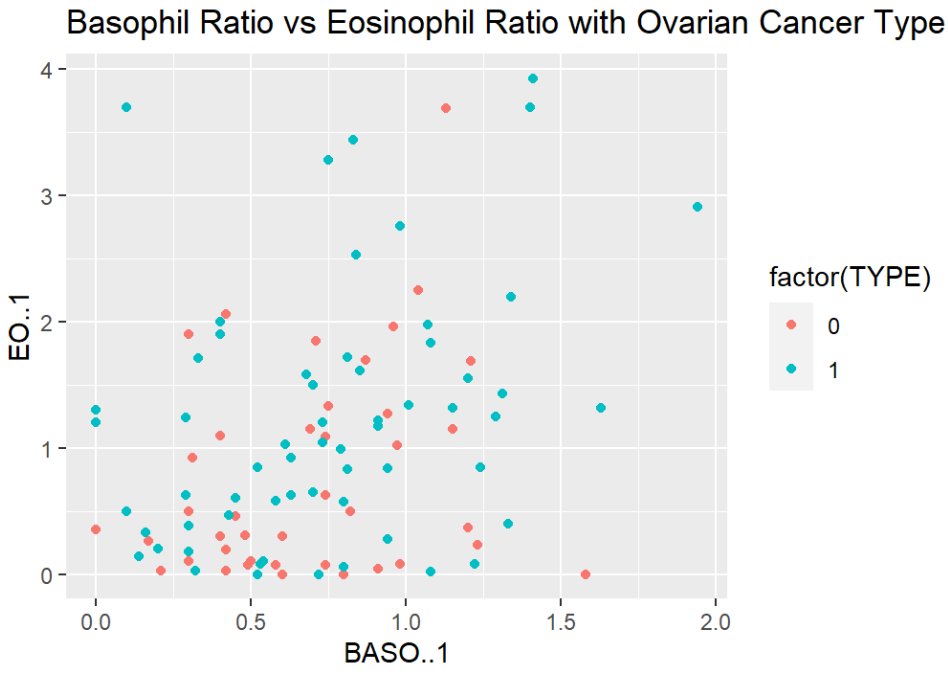






I will also look at a scatter plot of lymphocyte ratio and monocyte ratio with categorical labeling of the cancer type to determine how those biological components of cancer fluctuate with the cancer TYPE. Looking at the scatter plot, it can be visually clear how higher ratios of lymphocytes are prevalent

in patients with Benign Ovarian Cancer (TYPE = 1) compared to those with Ovarian Cancer (TYPE = 0). This shows that lymphocytes decrease with the progression of cancer. On the other hand, there seems to be no effect of monocytes on cancer and it appears stagnant with this plot.



Similar to the previous visual, I will also look at the eosinophil ratio and Basophil ratio along cancer TYPE since both of those biological chemicals are known to be high in patients with cancer or a terminating disease. Looking at the graph with just eosinophil and basophil, there does seem to be a relatively linear relationship between those two variables. This suggests that as eosinophil levels increase, so does basophil. But when looking at it with cancer TYPE, it can be concluded that higher basophil levels do correlate to both cancer TYPEs, which suggests that merely having Ovarian Tumors will increase basophil levels.

III. APPLYING DATA MINING TECHNIQUES

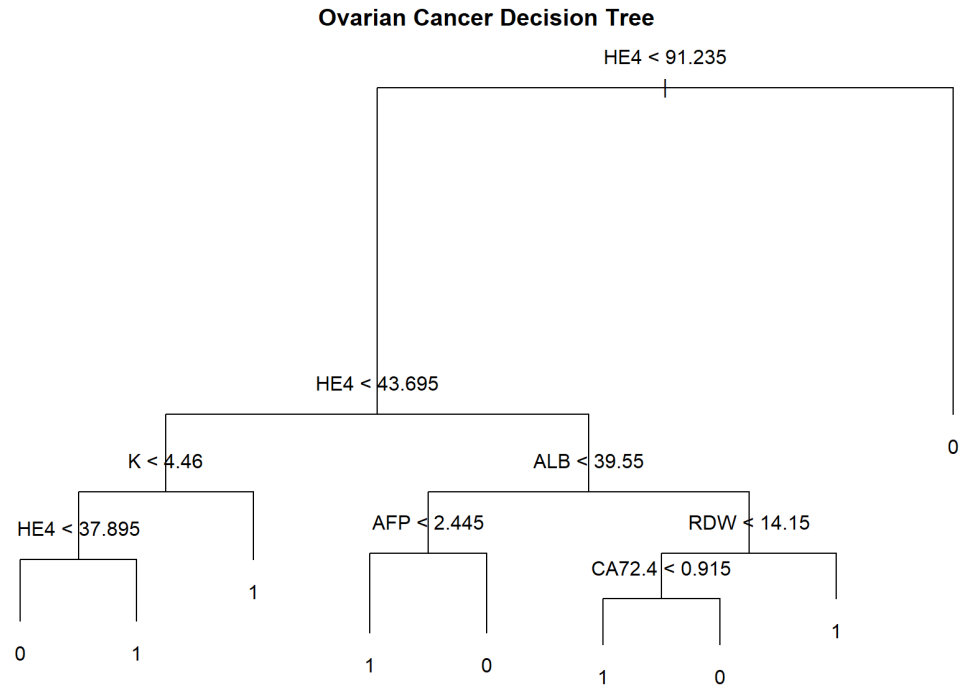
Now I'll begin applying data mining techniques to this dataset. I will use classification with decision trees to predict if a patient has Ovarian Cancer (TYPE = 0) or Benign Ovarian Tumors (TYPE = 1). I also apply clustering to analyze if there really are only two types of ovarian cancer present in this dataset. With this, I will have to make two copies of the cleaned dataset to use for both applications separately. The purpose of that will be to prevent any errors or disproportions in the analyses.

```
ovarian1 <- ovarian.data
ovarian2 <- ovarian.data
```

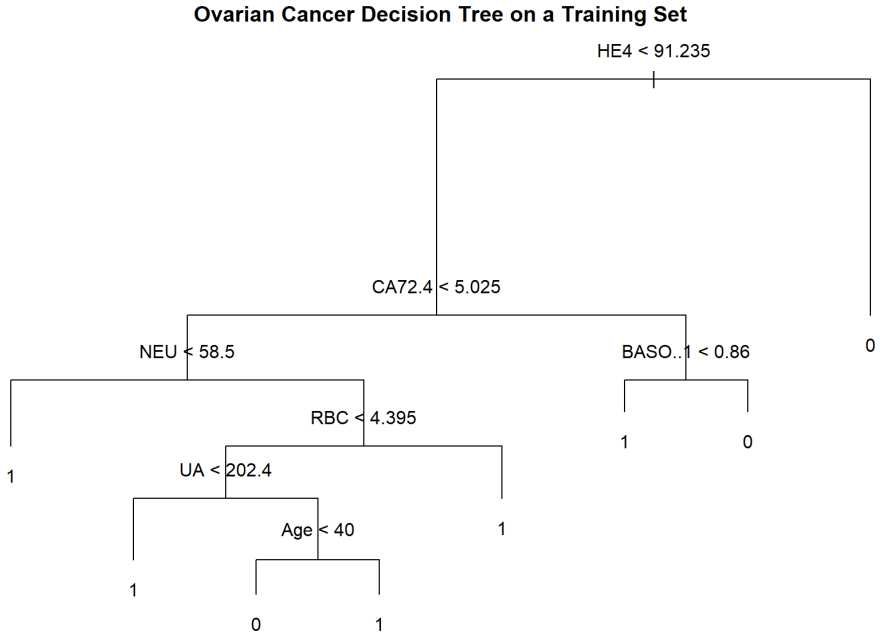
A-I. APPLYING CLASSIFICATION ANALYSIS TO PREDICT OVARIAN CANCER

I will use 'ovarian1' to build the decision tree. I will have to remove the variable SUBJECT_ID from this copied dataset it is not needed as it has zero correlations to the analysis of this classification. The variable TYPE will be the response variable for the decision tree with 0

= Ovarian Cancer and 1 = Benign Ovarian Tumor. After building the decision tree as is, we are left with 9 nodes and a misclassification error of 0.05882.



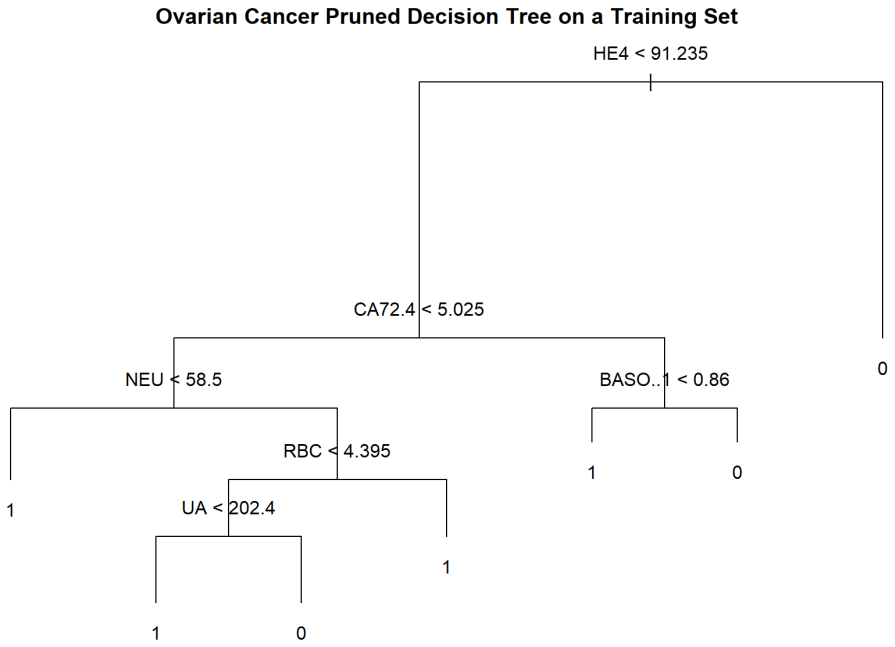
Now I will move on to training and testing this decision tree. I have split 'ovarian1' into a training set and a test set with the training set containing 80% of the data. Through the training, we can see that 8 nodes were used this time with a misclassification error of 0.06173. This is relatively very good.



Now I will test the decision tree on the test set and create a table of the results. From this table, it is calculated that the accuracy of the training model on the test set is roughly 71%. This suggests that the training model is an 'okay' model. It can be improved by pruning the model.

```
                  ovarian.test
dtree.ovarian.test 0 1
                 0 6 3
                 1 3 9
```

After pruning the decision tree, we are left with 7 nodes and a misclassification error of 0.07407. Despite the misclassification error continually increasing by 1%, it is still a low percentage and the number of nodes also decreased by 1. This is still a model that can be worked with.



Now the pruned model can be tested on the test set. After running it, we can see through the table of results that the accuracy has greatly improved with the accuracy increasing up to 95%. That's very impressive considering it began with 71%. This is a significantly better model to use to predict the outcome of ovarian cancer in a patient.

```
                  dtree.ovarian.pruned.test
dtree.ovarian.test  0  1
                 0  9  0
                 1  1 11
```

A-II. CLASSIFICATION ANALYSIS RESULTS AND DISCUSSION

With the first build of the model, there was a misclassification error of 0.05882 with 9 nodes. This suggests that the initial model was built very

```
Classification tree:
tree(formula = TYPE ~ ., data = ovarian1)
Variables actually used in tree construction:
[1] "HE4"    "K"      "ALB"    "AFP"    "RDW"    "CA72.4"
Number of terminal nodes:  9
Residual mean deviance:  0.2498 = 23.23 / 93
Misclassification error rate: 0.05882 = 6 / 102
```

accurately with only an error of 5% with a maximal node of 9 despite there being 49 other variables in the dataset. This is exceptionally good and if this dataset had more observations, it would be possible to further perfect this model.

Then, with the training set, the number of nodes decreases to 8 with a misclassification error of 0.06173. Even though the misclassification error increased by 1%, a misclassification error of 6% is still very good.

```
Classification tree:
tree(formula = TYPE ~ ., data = train.set)
Variables actually used in tree construction:
[1] "HE4"     "CA72.4"  "NEU"     "RBC"     "UA"      "Age"
[7] "BASO..1"
Number of terminal nodes:  8
Residual mean deviance:  0.2856 = 20.85 / 73
Misclassification error rate: 0.06173 = 5 / 81
```

The 1% misclassification error increase is also due to the fact the number of variables used in the model increased by one. In the initial decision tree, 6 variables (HE4, K, ALB, AFP, RDW, and CA72.4) were used, and then in the training set, 7 variables were used (HE4, CA72.4, NEU, RBC, UA, Age, and BASO..1). This suggests that a fluctuation of the number of nodes and number of variables can have an effect on the misclassification error. If more variables are used with fewer nodes, it could increase the misclassification error.

However, despite the low misclassification error, the accuracy of the training model was only 71%. This is probably due to the addition of that one variable. Even so, 71% accuracy is still relatively good considering the minimal amount of data that was available after the preprocessing.

The misclassification error increased by 1% again with it being 0.07407 after the pruning process. Although the number of variables did go back down to 6 (HE4, CA72.4, NEU, RBC, UA, and BASO..1), which has decreased the complexity should increase the accuracy of the pruned model. This proved to be true after running the pruned decision tree on the test set. The accuracy of the pruned model was 95% which is an extremely significant improvement from 71%.

```
Classification tree:
snip.tree(tree = dtree.ovarian, nodes = 37L)
Variables actually used in tree construction:
[1] "HE4"     "CA72.4"  "NEU"     "RBC"     "UA"      "BASO..1"
Number of terminal nodes:  7
Residual mean deviance:  0.3559 = 26.33 / 74
Misclassification error rate: 0.07407 = 6 / 81
```

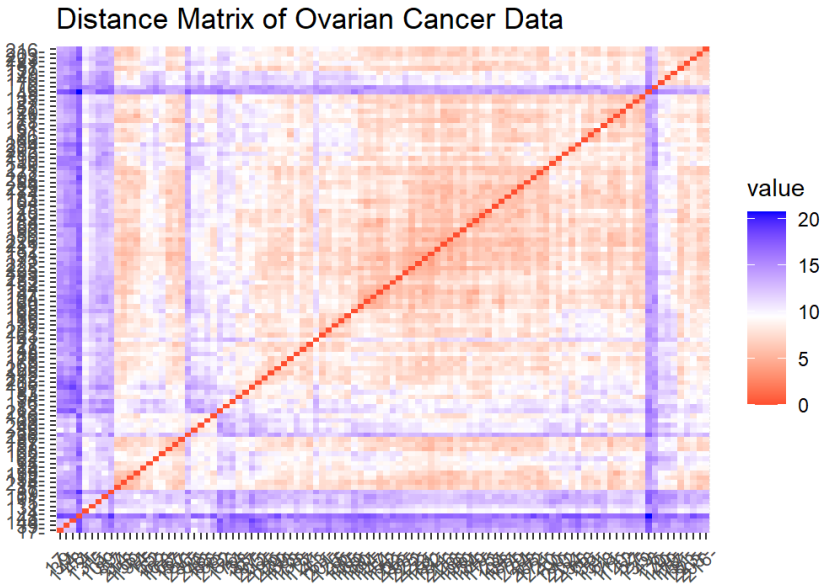
The accuracy of 95% of the pruned model suggests that this is as best as the model will get. A decision tree of 7 nodes and 6 variables is the best to predict ovarian cancer. This also means that 6 variables of HE4, CA72,4, NEU, RBC, UA, and BASO..1 contribute the most to ovarian cancer. The un-abbreviated names of these variables are human epididymis protein 4, Carbohydrate antigen 72-4, neutrophil ratio, Red blood cell count, urine acid, and Basophil Cell ratio, respectively, all of which do biologically affect or become affected by ovarian tumors. Also, the fact HE4 and CA72.4 consistently were the variables used in all decision trees suggests that human epididymis protein 4 and Carbohydrate antigen 72-4 are the two biological components that most greatly contribute to ovarian cancer.

B-I. APPLYING CLUSTER ANALYSIS TO CATEGORIZE OVARIAN TUMORS

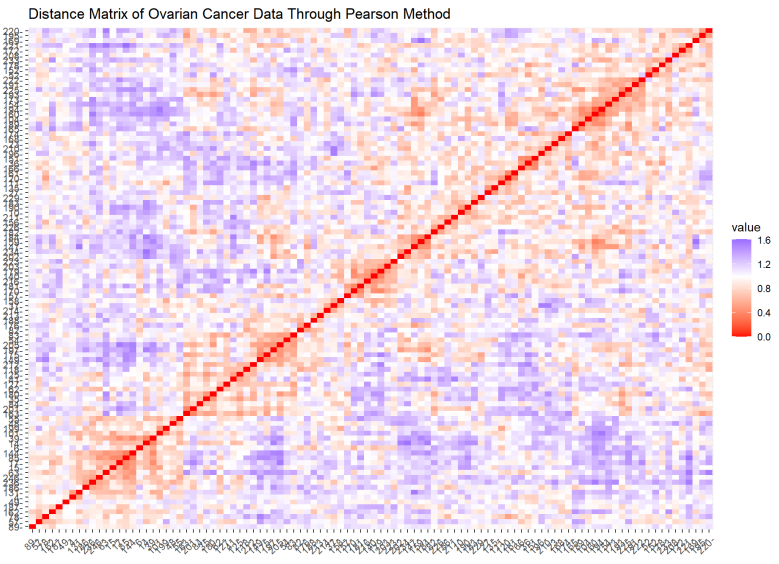
Here, I will use the copied dataset 'ovarian2' to conduct cluster analysis. This data mining technique will be used to categorize ovarian tumors. Before the clustering begins, I will also have to remove the variable SUBJECT_ID as it is not needed in this technique either. Additionally, I will have to change the variables Menopause and TYPE back into numeric in order to standardize all the variables as the variables need to be numeric in order to be standardized. I will use the summary() to check all variables are in the standard normal distribution.

```
> summary(ovarian2)
      AFP                 AG                Age          
 Min.   :-0.14852   Min.   :-2.11286   Min.   :-1.83662  
 1st Qu.:-0.12365   1st Qu.:-0.75949   1st Qu.:-0.80162  
 Median :-0.11196   Median : 0.03611   Median :-0.04262  
 Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000  
 3rd Qu.:-0.09094   3rd Qu.: 0.59675   3rd Qu.: 0.69913  
 Max.   : 9.98532   Max.   : 2.70082   Max.   : 2.23438  
      ALB                ALP                ALT          
 Min.   :-3.4939   Min.   :-1.7918   Min.   :-0.9141  
 1st Qu.:-0.3672   1st Qu.:-0.6169   1st Qu.:-0.5941  
 Median : 0.1420   Median :-0.1307   Median :-0.3234  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.7083   3rd Qu.: 0.3555   3rd Qu.: 0.1443
```

From here, I will inspect the data by looking at the distance matrix. Looking at the visual can be overwhelming with how many variables are being considered in the graph. This is usually the case for matrix visuals of this sort in healthcare research since there are always a lot of biological components to be looked at, especially in cancer and disease research. Also, simply looking at the graph by itself without considering the variables, it does appear there are only two types of ovarian cancers within this dataset.



I also looked at the distance matrix graph of the data using Pearson's method. The graph does appear to be cleaner, however, it is still



showing the same kind of expected number of clusters.

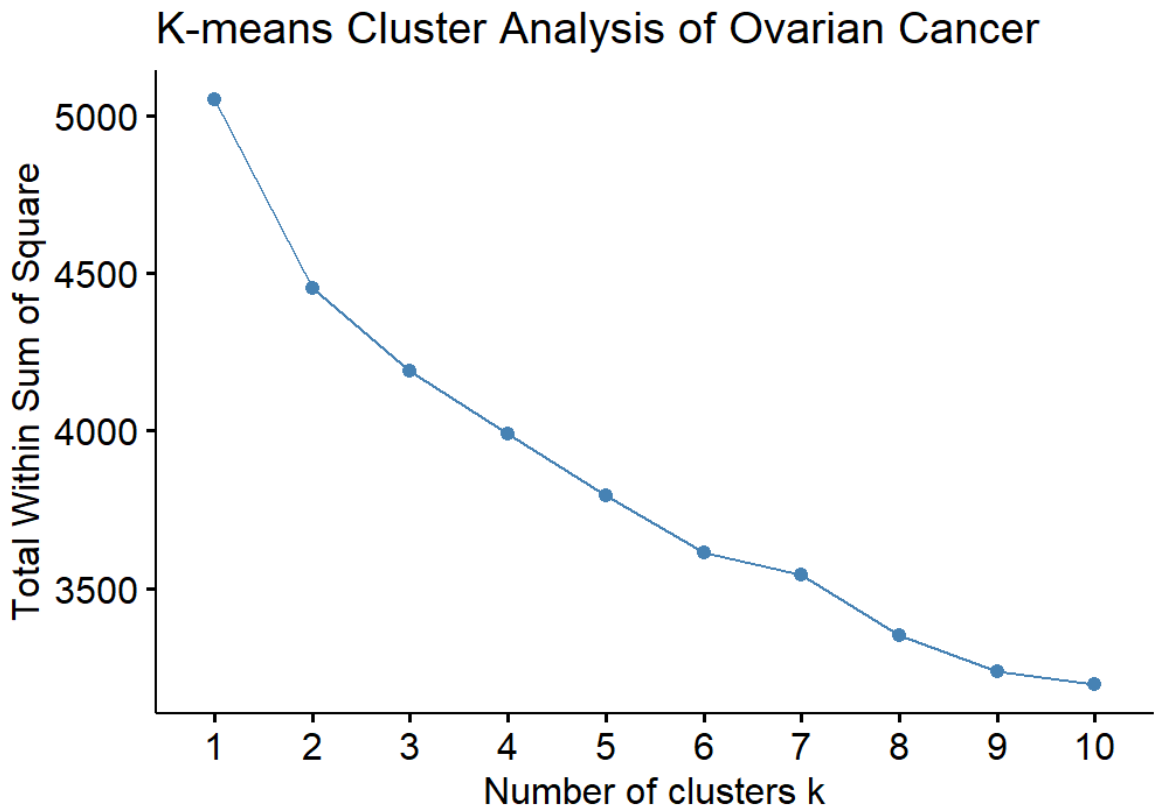
To begin, I started off with 3 centers to test if there really are more than just 2 types of ovarian cancer. The results gave me a 17.5% total variance in the cluster. This means only 17.5% of the data is explained through the total variance. This is pretty low and suggests that there are more clusters than I initially predicted.

```
Within cluster sum of squares by cluster:
[1] 1152.187 1146.727 1867.918
 (between_SS / total_SS =  17.5 %)
```

Then I decided to create a cluster analysis with 6 centers. I chose 6 to double up from the first cluster. That way I can have a clear understanding of the difference in analysis from both models. The result from the 6 clusters analysis showed to be better as the total variance increased to 24.6%, which is significantly better. This suggests that more than 3 clusters are needed to properly categorize the data.

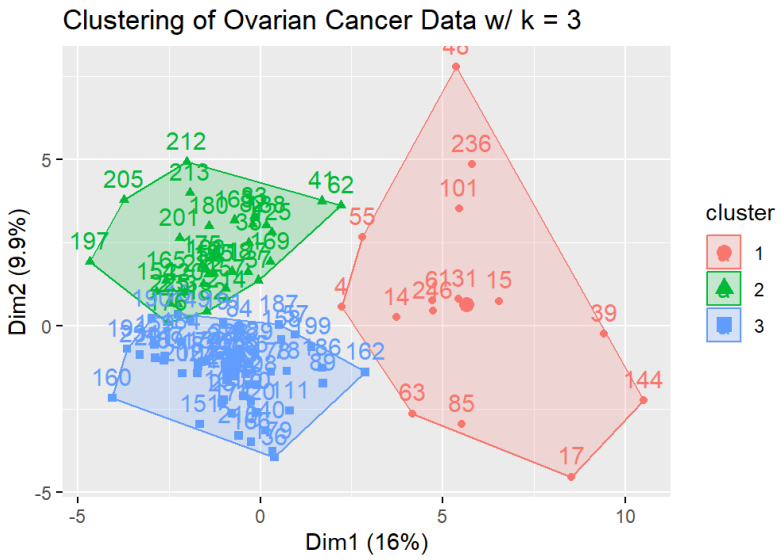
```
Within cluster sum of squares by cluster:
[1]    0.0000  768.7180 1018.6317  841.9057 1175.9519
 (between_SS / total_SS =  24.6 %)
```

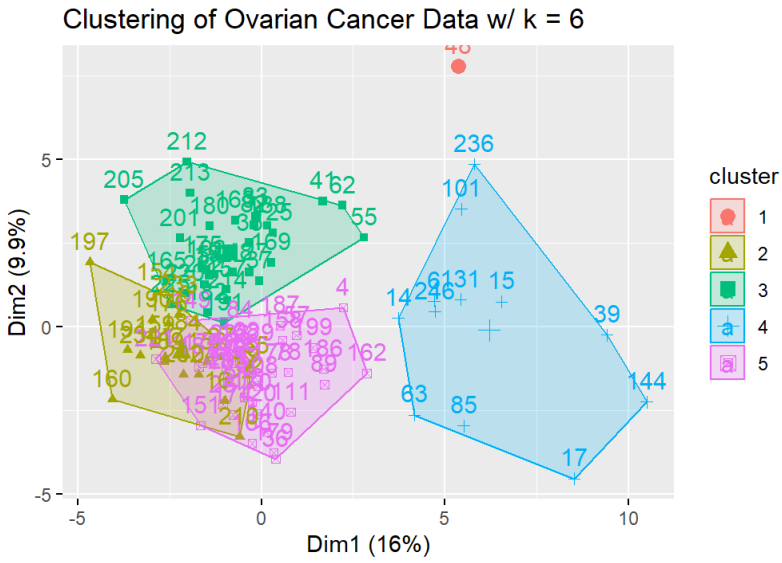
However, when looking at the K-means cluster analysis of the data in which we look at the number of clusters that can be explained by the data, it shows that fewer clusters can accurately represent the ovarian cancer data. Thus, there probably only are 2 types of ovarian cancers present within this particular dataset.



B-II. CLUSTER ANALYSIS RESULTS AND DISCUSSION

When the first cluster analysis was performed with 3 centers, the total variance was 17.5% and the one with 6 centers was 24.5%. This suggests that 6 clusters better represent the data. However, the visual graph of the clusters demonstrates that roughly 5 clusters properly categorize the ovarian tumors. There is also an outlier in the second graph, which shows that 5 clusters are more appropriate than 6.





However, since there is an overlap between the yellow and the pink clusters in the second graph, this suggests that 3 clusters are actually better.

Another reason why I believe 2 clusters might be better despite the low total variance is because of the alignment of the green and blue clusters in the first graph. This tells me that perhaps the green and blue should all be one cluster. Likewise, in the second graph, the green, yellow, and pink clusters are in the same alignment and area. There's also the cluster to the right in both graphs that is relatively the same size and stayed in the same location. This also suggests that perhaps 2 clusters are the most appropriate for categorizing these ovarian tumors.

IV. SUMMARY

In conclusion, decision trees are a great way of both visualizing and conducting prediction methods for ovarian cancer. The accuracy of the final tree model was 95%, which again, is significantly good for such an analysis. Decision trees are a great tool for healthcare research to get direct and clear results of important health concerns and questions.

The cluster analysis also showed that there is much more to do to properly analyze and categorize ovarian cancers and tumors. The low total variances in both clusters suggest that perhaps more variables need to be recorded and analyzed in order to properly categorize these ovarian tumors. I also am making the assumption that perhaps the data needed to be standardized in a different way.

All in all, ovarian cancer needs to be researched more. Conducting further data mining research and collecting more data to create better diagnostics for early detection is imperative to the health of these patients.